MEGABRAIN

A Unified Knowledge Architecture for the Modern Researcher

Author: Sangram Gayal | February 2026 | **(BRS Summary)**

The Problem: Knowledge Management Crisis

Modern researchers face a paradox: information has never been more abundant, yet actionable knowledge remains frustratingly elusive. Over 2.5 million scholarly articles are published annually, yet researchers spend 19% of their time simply re-locating information they have already encountered.

Information Avalanche

2.5M+ new papers per year across 30,000+ journals. The corpus doubles every 18 years.

Tool Fragmentation

Researchers juggle 4-7 disconnected tools: Zotero, Obsidian, NotebookLM, Notion, Evernote.

Knowledge Decay

Insights from previous reading disappear. 1 full day per week lost to re-finding information.

What researchers actually need:

Unified semantic search across an entire academic library • Methodological preferences that inform interpretation • Connections between formal research and adjacent insights • Data sovereignty with local-first operation • Institutional-grade tools at personal economics (< $1/month)

The Solution: Megabrain

Megabrain is an AI-native, local-first knowledge architecture that unifies three distinct knowledge domains into a single semantically searchable system powered by vector embeddings and retrieval-augmented generation (RAG).

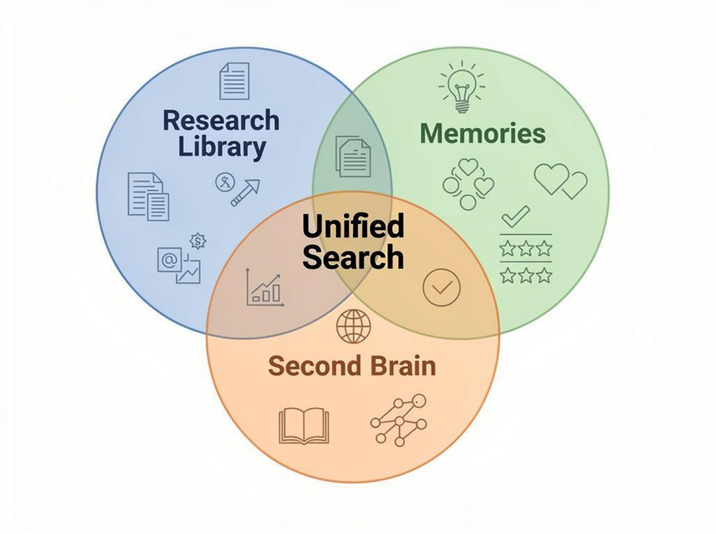

Figure 1: Three contextually separated vector indexes unified through semantic search

Index 1 - Research Library

Auto-synced from Zotero. Full-text PDF extraction, 1,000-character semantic chunking, APA citation metadata. Handles 1,000+ papers.

Index 2 - Memories

Captures tacit knowledge: methodological preferences, theoretical stances, domain expertise. Shapes how research is interpreted.

Index 3 - Second Brain

Web articles, YouTube transcripts, podcasts, tweets, and industry reports. Source of cross-pollination and serendipitous discovery.

Technical Architecture & Data Flow

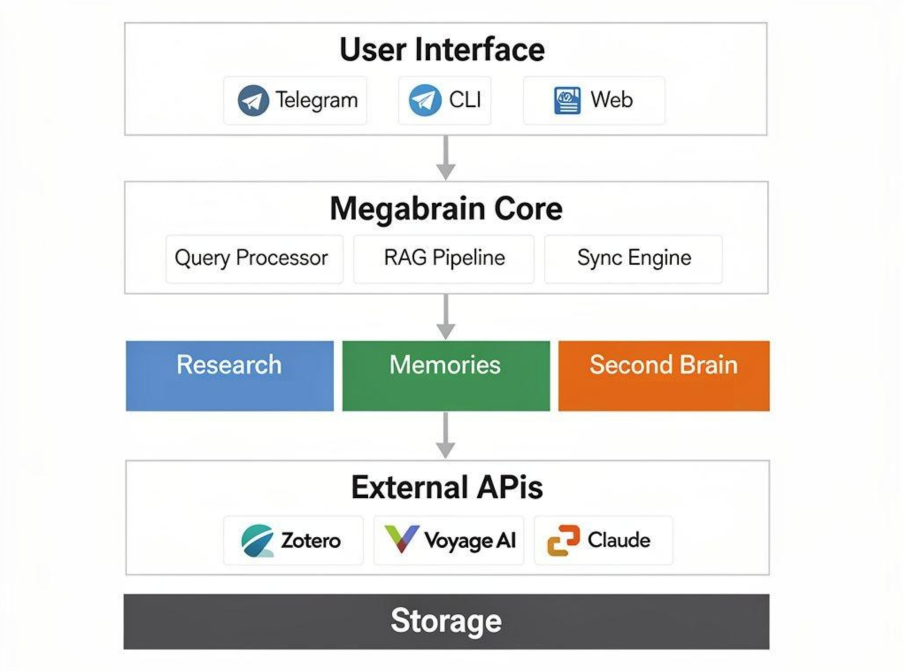

Figure 2: End-to-end ingestion and query/synthesis pipeline

Technology Stack

Component	Technology	Role
Vector DB	FAISS (local)	High-performance nearest-neighbor search. Runs entirely on researcher's machine. No server required. Handles 5,000+ papers (<5 GB).
Embeddings	Voyage AI voyage-3	1,024-dimensional vectors optimized for academic retrieval. ~$0.10 per 1M tokens; full library indexing < $5 one-time cost.
Reranking	Voyage rerank-2	Two-stage retrieval: vector similarity (top 20) followed by cross-attention rescoring (top 8). +15-20% relevance improvement.
Language Model	Claude 3.5 Sonnet	RAG synthesis with precise citation formatting (APA/MLA/Chicago) and near-zero hallucination on grounded context.
Orchestration	Python + LangChain	Custom multi-index retrieval, contextual weighting, and Zotero incremental sync every 30 minutes.
PDF Ingestion	PyMuPDF + pyzotero	1,000-character semantic chunking with 200-char overlap. Auto-fallback to abstract if PDF unavailable (paywalled).

Key Benefits

Speed Research Velocity

76% reduction in literature review time. 8.5 hrs -> 2 hrs per topic. Natural language queries return synthesized, cited answers in < 3 seconds.

Privacy Cognitive Sovereignty

Fully local-first. Papers, embeddings & indexes on researcher's own machine. Optional cloud augmentation. No vendor lock-in.

Quality Knowledge Integrity

Three separate vector indexes prevent contamination between peer-reviewed findings, personal insights, and general reading.

Cost Economic Efficiency

< $5 setup, < $1/month operating. Replaces a $379-546/year tool stack (Mendeley, Notion, Elicit, Readwise) at 67% lower cost.

How It Works: End-to-End Flow

1 **Ingest**

Zotero auto-sync every 30 min. PDF text extracted, chunked into 1,000-char segments, embedded via Voyage AI.

2 **Store**

Vectors stored in three contextually separate FAISS indexes. All data remains local on researcher's machine.

3 **Retrieve & Synthesize**

Natural language query -> embed -> FAISS search (top 20) -> Rerank (top 8) -> Claude RAG synthesis -> cited response.

Development Roadmap Phase 1 (M1-2): Zotero sync + RAG core • Phase 2 (M3-4): Memories & Second Brain • Phase 3 (M5-7): Obsidian plugin, browser extension • Phase 4 (M8-10): Proactive alerts & topic clustering • Phase 5 (M11-15): Team collaboration & mobile app